Estimation For The Beta Coefficients For The Multiple Regression

علي لطيف عارف
كلية العلوم–جامعة ذي قار

الخلاصة

في العديد من دراسات الانحدار,هناك طموح لمقارنة الأهمية النسبية للمتغيرات التفسيرية حيث هناك مقاييس مختلفة يمكن أن تستعمل .مثل قيم t , معاملات الانحدار , وقيم p وغيرها والأكثر استعمالا هي طريقة معاملا ت الانحدار القياسية (معاملات بيتا) والتي يمكن أن تحسب بطريقتين وكلاهما يؤدي إلى نفس النتيجة .لذا من الصعب في اغلب الأحيان معرفة أي من المتغيرات التوضيحية الأكثر أهمية في تحديد المتغير المعتمد حيث إن القيمة التي تعتمد عليها معاملات الانحدار هي في اختيار وحدات القياس إلى X .كما تناول البحث كيفية حساب الانحراف المعياري الجزئي في حالة المتغيرات المستقلة التي لها توزيع مشترك غير معرف كذلك تناول البحث العلاقة بين المعاملات الجديدة والتوصل إلى أن لها نفس النسبة بين قيم t والجزء المهم أيضا كان بيان حالة الضعف الوحيدة عند استخدام المعاملات . الجانب التطبيقي في هذا البحث تناول البيانات الخاصة بسوق الأسهم المالية حيث سعر السهم لكل حصة أو سهم يمثل المتغير المعتمد Y , ومتغيران مستقلان معدل العائدات x_1 والنسبة السنوية من الحصة x_2 البرنامج الإحصائي المستخدم في تحليل البيانات هو (SPSS) .

Abstract

In many regression studies , there is an ambition to compare the relative importance of the explanatory variables where there are different measures that can be used : t values , regression coefficient , p values , and the frequently used is standardized regression coefficients (beta coefficients). It is can be calculated in two ways , with both leading to the same result ,therefore , it is often difficult to say which of the explanatory variables is most important in determining the value of the dependent variable , since the value of the regression coefficients depends on the choice of units to measure X. The research is dealing to calculate partial standard deviation when the joint distribution is unknown , also the research explain the relationship between the new coefficients will be similar to comparing t values, and the important part is to explain the one weakness of the standardized coefficients when the coefficients are used . As the partial (applied) side of the this research has conserving the special data of the Stock market , the price per share Y , two variables that are thought to influence stock price average equity X_1, and annual rate of dividend X_2 . a statistical programmed is here used that is (SPSS,) for the analysis of the data.

Introduction

When regression models are used , there are different measures that can be used to compare the effect of the explanatory variables ,It is often difficult to say which of X variables is most important in determining the value of the dependent variable , since the value of the regression coefficients depends on the choice of units to measure X . Standardized regression coefficients (beta coefficient) are frequently used in quantitative social sciences. they are used for many purposes :selecting variables ,determining the relative importance of explanatory variables ,comparing the effect of changing different variables. Standardized coefficient still used when they have been so severely criticized? One reason is that the question of relative importance of different variables is very difficult to answer, and to used standardized coefficient seems to be an easy solution. Another reason might be that none of the mentioned critiques .Recall that the size of the coefficient depends in part on the mean and variance of the independent variable. Often, our independent variables have widely varying mean and variances , meaning that the resulting coefficients can't be directly compared. the usual way this is done in the statistics (not just in regression)is to convert the values of the independent variable into *deviations* .Deviations are very similar to a Z score.

Standardized regression coefficients

Consider an estimated regression equation of y on $x_1, x_2, \ldots\ldots., x_k$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots + \beta_k x_k \qquad \ldots\ldots.(1)$$

The usual interpretation of a regression coefficient β_i is the average change in y when x_i is change by one unit, given that the other variables are held fixed. The regression coefficient is sometimes called *partial regression coefficients* to stress the fact that the estimated effect of changing a variable is conditional on the other variables in the regression equation held constant. For this interpretation to be valid, assume that the model is correctly specified and that the included variables appear only once in the equation (when the square of a variables is included to model nonlinear relationships or interactions are included, the interpretation becomes more difficult) .Another important remark about the usual interpretation about the regression coefficient concerns causation. When the data do not come from an experiment, the regression coefficient is only a descriptive characteristic of the sample. It might not be correct to draw any conclusion about what will happen if x_i is change .A common modeling objective is to be able to compare regression coefficients with respect to size. This comparison is difficult when the variables are measured in different units. To overcome this problem, we used standardized coefficient ,arguing that standardized variables are measured in comparable units, standardized regression coefficients can be calculated in two ways , with both leading to the same result . The idea is to express values for all independent variables in a dataset in terms of the number of standard deviations that variable is from the its mean. Procedurally, for each variable we should first calculate the mean and variance across all observations. then for each observation, we subtract out the mean and divide by the variance for each variable. So ,

$$\frac{y-\bar{y}}{s_y} = \beta_1^* \frac{x_{1i}-\bar{x}_1}{s_{x_1}} + \beta_2^* \frac{x_{2i}-\bar{x}_2}{s_{x_2}} + \ldots..\beta_n^* \frac{x_{ni}-\bar{x}_n}{s_{x_n}}{}_i \qquad \ldots(2)$$

$$x_i^* = \frac{x_i-\bar{x}_i}{s_i} \qquad i=1,2,\ldots\ldots.,k \qquad \ldots\ldots..(3)$$

$$y^* = \frac{y-\bar{y}}{s_y} \qquad \ldots\ldots\ldots..(4)$$

Where $\bar{x}_i$ and $\bar{y}$ are the means of each explanatory and dependent variables respectively in the sample and s_i , s_y are the standard deviations, and then estimate (1) by (OLS) with standardized variables each "standardized " variable has a mean of (0) and variance of (1).

The regression coefficients in this equation are the standardized regression coefficients. This standardization is sometimes recommended to improve computational accuracy , The standardized coefficients have slightly different meaning . if $\beta_1^* = 0.55$ then a 1 standard deviation change in x_1 results in 0.55 standard deviation increase in the dependent variable.

Another possibility is first to calculate the regression coefficients by using unstandardized variables and then multiply them by the ratio between the standard deviation of the respective independent variable and standard deviation of the dependent variable :

$$\beta_i = \hat{\beta}_i (S_i / S_y) \qquad \ldots\ldots\ldots\ldots\ldots(5)$$

Where β_i is the standardized regression coefficient . The β_i in (4) are identical to the standardized coefficient estimates from OLS on the standardized variables in (3,4)
Another formula shows estimates the standardized coefficients by using correlation coefficients as :

$$\beta_1^* = \frac{r_{y1} - (r_{12} * r_{y2})}{1 - r^2_{12}} \qquad \ldots\ldots(6)$$

$$\beta_2^* = \frac{r_{y2} - (r_{12} * r_{y1})}{1 - r^2_{12}} \qquad \ldots\ldots(7)..$$

The regression coefficient between two standardized variables is equal to the covariance of them. Therefore the correlation matrix is the same as the covariance matrix. that is the variances of the standardized variables equal to 1 , and the covariance equal the correlations ,also by the use of standardized coefficients you can compute R^2 by :

$$R^2 = \sum \beta_k^* * r_{yk} \qquad \text{OR}$$

$$R^2 = (\beta_1^*)^2 + (\beta_2^*)^2 + 2 * \beta_1^* * \beta_2^* * r_{12}$$

Inconsistency in the calculation of standardized coefficients

Consider the following example ;blood pressure is regressed on weight and height
blood pressure= $\alpha + \beta_1 weight + \beta_2 height + \varepsilon.$ (8)

the standardized coefficient is calculated by multiplying $\hat{\beta}_1$ by s_1 The inconsistency in this calculation arises because $\hat{\beta}_1$ and s_1 refer to different populations. The regression coefficient , $\hat{\beta}_1$, is interpretable only under the restriction that height is held constant ,whereas s_1 measures the spread in weight among all people in the sample , regardless of their heights .to clarify this consider an example .

Table: (1) Blood pressure ,Height ,and Weight for nine patients

Group	Blood pressure (mm Hg)	Height Centimeters)	Weight (Kilograms)	Standard deviation (S^*) of weight within the group
A	103	160	55	
A	123	160	60	5
A	112	160	65	
B	146	170	65	
B	131	170	70	5
B	119	170	75	
C	152	180	75	
C	133	180	80	5
C	140	180	85	

In this table keeping height constant is equivalent to be longing to one of groups A,B,C when the standardized coefficient is calculated , $\hat{\beta}_i$ is multiplied with the standard deviation in the whole sample (s=9.7).
but this estimate of 9.7 is irrelevant when the analysis is conditioned on height held constant because the conditional variation in weight is much smaller than 9.7 ($\overset{*}{s} = 5$ in each group)

Calculation of partial standard deviation

If the independent variables follow a multivariate normal distribution, then it is easy to calculate partial standard deviation . In most applications ,the joint distribution is unknown , and it is therefore not possible to calculate the conditional distribution. The conditional standard deviation for x_i ,can be estimated by regressing x_i on the other independent variables. Under an assumption of homoskedasticity ,the partial standard deviation can be estimated by the square root of the mean squared error. Alternatively, this estimate could easier be obtained by using the variance inflation factor (VIF) computed by most statistical computer packages .When y is regressed on $x_1, x_2, ..., x_k$ each independent variable is associated with a VIF

$$VIF = 1/(1 - R^2{}_{k-1}) \qquad ....(9)$$

Where ($R^2{}_{k-1}$) is the coefficient of determination when x_i is regressed on the (k-1) other independent variables .The VIF tells us :
The degree to which the standard error of the predictor is increased due to the predictor's correlation with the other predictors in the model.
Then the partial standard deviation is

$$S^*{}_i = \frac{S_i}{\sqrt{VIF_i}} \sqrt{n-1/n-k} \qquad ......(10)$$

the partial standard deviation will remain a more relevant indicator of spread than unconditional standard deviation . If all variables are standardized first, using the partial standard deviation gives

$$x_i^* = \frac{x_i - \bar{x}_i}{s_i^*} \qquad .....(11)$$

And the new standardized coefficients will be equal to the ordinary regression coefficients , calculated on the standardized variables.

The new standardized coefficient

Assume we want to reduce the variance in y by manipulating one and only one of the independent variables . all of the other variables have the remain unchanged. Which variable should we change? There are at least two factors relevant for this decision :
1) What is the effect of changing a variable given the other held constant?
2) By how much is it possible to change a variable without changing the other independent variables?

The regression coefficient $\hat{\beta}_1$ answers the first question and the answer of the second question when is used the partial standard deviation S_i^* , the relative importance should be evaluated by estimating the expected change in y caused by the greatest likely change in x_i , for example blood pressure is regressed on weight and height , the variation in weight among people of the same height is much better indicated by the partial standard deviation than by the standard deviation . We can not expect people of the certain height to have the same variation in weight as people with different heights.

Relationship between standardized coefficients and *t* values

In most regression applications, a ***t*** value is calculated for each independent variable. These ***t*** values are often used as a criteria and for tests significance . the t values are

$$t_i = \frac{b_i}{s_{b_1}} \qquad \text{...... ..(12)}$$

where

$$s_{b_1} = \frac{s_e}{\sqrt{\sum (x_{ai} - \bar{x}_i)^2 (1 - r_i^2)}} \qquad i=1,2,\ldots,k \qquad \text{.....(13)}$$

And s_e is the standard deviation of the residuals.

Comparing the magnitude of two t values by calculating the ratio between them gives

$$\frac{t_1}{t_2} = \frac{b_1 \sqrt{\sum (x_{a1} - \bar{x}_1)^2} \sqrt{1 - r_1^2}}{b_2 \sqrt{\sum (x_{a2} - \bar{x}_2)^2} \sqrt{1 - r_2^2}} \qquad \text{......(14)}$$

The ratio between the two variables β^* values is

$$\frac{\beta_1^*}{\beta_2^*} = \frac{b_1 (\sqrt{\sum (x_{a1} - \bar{x}_1)^2} / (\sqrt{n-k}) \sqrt{1 - r_1^2}}{b_2 (\sqrt{\sum (x_{a2} - \bar{x}_2)^2} / \sqrt{n-k}) \sqrt{1 - r_2^2}} \qquad \text{........(15)}$$

$$= \frac{b_1 \sqrt{\sum (x_{a1} - \bar{x}_1)^2} \sqrt{1 - r_1^2}}{b_2 \sqrt{\sum (x_{a2} - \bar{x}_2)^2} \sqrt{1 - r_2^2}} \qquad \text{............(16)}$$

From which it is seen that

$$\frac{\beta_1^*}{\beta_2^*} = \frac{t_1}{t_2} \qquad \text{.....(17)}$$

Hence , instead of comparing the new standardized coefficients, we could compare ***t*** value

Application

Stock market analysts are continually searching for reliable predictors of stock price. Consider the problem of modeling the price per share, Y, of electric utility stocks. Two variables that are thought to influence stock price are return on average equity x_1 , and annual rate of dividend x_2 . the stock prices, returns on equity, and dividend rates for a sample of 16 nonnuclear electric utility stocks are shown in the accompanying table . the model is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Table (2) : Nonnuclear stocks

Y	X_1	X_2
25	15.2	2.60
20	13.9	2.14
15	15.8	1.52
34	12.8	3.12
20	6.9	2.48
33	14.6	3.08
28	15.4	2.92
30	17.3	2.76
23	13.7	2.36
24	12.7	2.36
25	15.3	2.56
26	15.2	2.80
26	12.0	2.72
20	15.3	1.92
20	13.7	1.92
13	13.3	1.60

Source: United Business Investment Report

Below is the regression output :

Table (3) : Values of Beta regression coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-9.954	3.405		-2.923	.012
	x1	.476	.186	.190	2.556	.024
	x2	11.194	.877	.947	12.761	.000

a. Dependent Variable: y

Table (4) : **Correlation Matrix**

		y	x1	x2
Correlation	y	1.000	.178	.945
	x1	.178	1.000	-.012
	x2	.945	-.012	1.000

From table (3) the regression equation is :

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$= -9.954 + 0.476\, X_1 + 11.194\, X_2$$

$$\beta_1^* = \hat{\beta}_1 \frac{S_{x1}}{S_y}$$

$=0.476 * (2.3221 / 5.830)$

$=0.19$

$$\beta_2^* = \hat{\beta}_2 \frac{S_{x2}}{S_y}$$

$=11.194 * (0.493 / 5.830)$

$=0.947$

OR from table (4) and by using the formula (14 , 15) we compute the standardized coefficients which given the same result .

β_1^{*} =0.178 – (-0.012 * 0.945) / 1 – (-0.012)^2
=0.19
β_2^{*} =0.945 – (-0.012 * 0.178) / 1 – (-0.012)^2
=0.947

Computing R^2

R^2 =(0.19 * 0.178) + (0.94 * 0.945)
=0.0338 + 0.8949 = 0.92 OR

R2 = $(0.19)^2$ + $(0.94)^2$ + 2*(0.19) *(0.94) * (-0.012)
=0.0361 + 0.8836 – 0.0042
= 0.9297 – 0.0042
= 0.92

t_1 / t_2 = 2.556 / 12.761 = 0.202

β_1^{*} / β_2^{*} =0.19 / 0.94 = 0.202

Hence standardized coefficients tell you how increases in the dependent variables affect relative position within the group . You can determine whether one standard deviation change in one independent variable produces more of a change in relative position than one standard deviation change in another independent variable.

Conclusions

1-Standardization is the process by which the new data transformed into new variables that have mean 0 and variance of 1
2- Standardized variables don't have any units, they are just numbers you can compare them regardless of the scale of the original variables.
3- The standardization should be done with partial standard deviations instead ordinary standard deviation .
4- Comparing new standardized regression coefficients is equivalent to comparing ***t*** values
5- Standardized coefficients provide an easy means for computing R^2 .

References

Afifi, A.A, and Clark ,V.(1990) "Computer Aided multivariate analysis (2nd ed), New York :Van Nostrand Reinhold

Bring, j. (1993),"Variable importance by partitioning R^2 "(research report 1993:9),Uppsala University ,Dept of statistics

Greenland ,S,Schlesselman j.j and Criqui, M,H. (1986),"the fallacy of employing standardized regression coefficient and correlations as measure of effect" American journal of epidemiology,123,203-208.

Karl,G,Joreskog .(1999),"How large can a standardized coefficient be "

Kim ,j .o ,and ferree,G.(1981)"Standardized in causal analysis" sociological methods and research , 1981,10,187-210

Mayer,L.S,and Younger ,M.S(1976),"Estimation of standardized coefficient " journal of the American statistical association , 1976,71,154-157

Mosteller ,F,and Tukey,j.W,(1977) " Data analysis and regression " New York: Addison –Wesley